Harnessing Interdisciplinarity to Promote the Ethical Design of AI Systems

Harnessing Interdisciplinarity to Promote the Ethical Design of AI Systems

Menisha Patel[1], Helena Webb[1], Marina Jirotka[1], Alan Davoust[2], Ross Gales[1], Michael Rovatsos[3], Ansgar Koene[4]
[1]University of Oxford, Oxford, United Kingdom
[2]Université du Québec en Outaouais, Gatineau, Canada
[3]University of Edinburgh, United Kingdom
[4]University of Nottingham, United Kingdom
menisha.patel@cs.ox.ac.uk
helena.webb@cs.ox.ac.uk
marina.jirotka@cs.ox.ac.uk
ross.gales@cs.ox.ac.uk
alan.davoust@uqo.ca
mrovatso@inf.ed.ac.uk
ansgar.koene@nottingham.ac.uk

Abstract: There is a growing global awareness that increasingly powerful AI technologies are being developed which have the potential to reshape societies and institutions. ICT researchers and practitioners are under pressure to consider and reflect on the motivations, purposes and possible consequences of their innovations. Whilst it has long been recognised that technological innovations have social and ethical impacts, a gap remains in practice between ethics and social science research on the one hand, and computer science and engineering on the other. Few opportunities exist to incorporate ethical or social reflection into system development in order to design more responsible technologies. We argue that interdisciplinarity is fundamental to identifying pathways to best practice in the design and development of AI innovations - including their deployment in, and impact on, society. In this paper, we detail our experience of conducting an `ethical hackathon' as a tool for the facilitation of the ethical design of AI systems. This non-conventional hackathon model draws on Responsible Innovation (RI) and places primacy on the the integration of ethics by bringing together a range of disciplines as a necessary part of addressing a design task. In an ethical hackathon, computer scientists and engineers collaborate closely with specialists from other fields in order to learn how to work together effectively to design more responsible technologies. Teams which include computer scientists, engineers, ethicists, social scientists and business students, complete a task that requires them to anticipate and reflect on the social and ethical issues that may emerge from an innovation, and also consider how to address these in their technical designs. Through a qualitative analysis we highlight the significant potential of the model to facilitate the ethical design and development of AI systems. However, we also identify several barriers to the success of the approach and conclude that in order to conduct a successful ethical hackathon, and engender a truly interdisciplinary consideration of the ethics of AI, careful design and management of participants' expectations is required. To this end, we conclude the paper by providing design implications which build on our experiences.

Keywords: interdisciplinarity, ethics, hackathons, responsible innovation

1. Introduction

There is a growing awareness that increasingly powerful technologies such as AI are being developed which have the potential to reshape societies and institutions. ICT researchers and practitioners are under increasing pressure to consider the motivations, purposes and possible consequences of their innovations. Whilst it has long been recognised that technological innovations have social and ethical impacts, there is an awareness that the current approach to ethics processes and procedures within ICT may not be adequate for contemporary research. A gap remains in practice between ethics and social science research on the one hand, and computer science and engineering on the other. Despite a rich history of reflecting on the ethics of ICT in the field of computer ethics (Johnson & Miller 2009, Floridi 2010), the development of codes and standards in ICT (ACM 2018, IEEE 2018, Society 2015) and calls from funding councils for interdisciplinary research, both groups have tended to work independently in ICT development- whilst acknowledging the importance of each other's work. Few opportunities exist to incorporate ethical or social reflection into system development to design more responsible technologies. This lack of impact across disciplinary boundaries has long been recognised as problematic.

AI ethics is fundamentally interdisciplinary. Identifying pathways to best practice in the development of AI innovations, including their impact on society requires the engagement of experts and stakeholders from a broad range of fields. The importance of interdisciplinarity in AI ethics is recognised by the AAAI and ACM as well as industry organisations, and is driven forward through conference and workshop events (Albrecht et al. 2015, AI Matters 2017) and the development of practical frameworks and guidelines (ACM 2018, Cutler et al. 2018). However, the existence of different disciplinary boundaries and sometimes competing understandings (Edmondson & Harvey 2018) makes meaningful interdisciplinary engagement difficult to achieve. When such interdisciplinary engagement relates to complex ethical issues, it becomes harder still. In this paper, we discuss our work that has adapted the notion of the hackathon to develop a tool to promote interdisciplinarity in the ethical design of AI systems. We describe how we have taken the conventional hackathon model and reshaped it through the lens of Responsible Innovation (RI) (which will be detailed later in the paper) into an interdisciplinary event that gives equal importance to technical and ethical issues of design. We reflect on the process and outcomes of a recent event using this model to highlight the opportunities and challenges presented by our `ethical hackathon'. We highlight the significant potential of the model to facilitate interdisciplinary collaboration in the ethical design of AI systems. However, we also identify potential barriers to success in relation to: the framing of ethics during the event; engendering teamwork and collaboration; and varying participant familiarity with the hackathon model. Each of these barriers is built in some way on disciplinary variations in expertise and expectations amongst event participants, demonstrating the particular challenges that exist in successfully forefronting interdisciplinarity in AI ethics.

1.1 Background

The traditional hackathon model is increasingly being appropriated to address complex issues in the contemporary innovation landscape. Given this, it is gaining recognition as an important mechanism through which to facilitate the exploration and resolution of a diverse array of cross-disciplinary innovation issues. For example, hackathons have been held to explore the intersection of technology and health (Birbeck et al. 2017); to support communities in developing their own civic technologies (Taylor et al.2018); and even within journalism (Boyles 2017). The non-conventional hackathon format, or the *mainstreaming* of hackathons (Taylor & Clarke 2018), has been characterised in various ways- as a growing phenomenon with the potential to bring together multi-disciplinary stakeholders in co-creation, open design or participatory design. The value of this approach is seen in the provision of a mechanism through which we can harness relevant expertise and perspectives to address the complexities of contemporary innovation.

Traditional hackathons have a well-defined scope. It has been argued that newer forms generally lack this clarity (Porter et al. 2017, Bir-beck et al. 2017). Indeed, the traditional hackathon model has been extended to the extent that it may involve no technical element at all. As a result, questions have arisen concerning the scope of these extended forms and how they will function in practice. This particularly relates to how they should be structured to support the goals of their use in non-conventional and interdisciplinary contexts (Porter et al. 2017). Additionally, there have been concerns regarding what the impact of these new forms should be (Briscoe 2014) - for prototyping or building capacity etc., and how such impacts should be reached. One of the core aspects of the newer forms of hackathon - interdisciplinarity - has also come under scrutiny. This is because although the popularity of such `cross-boundary working' (Edmondson & Harvey 2018) has been growing in different areas, it has not been without its challenges. It has often been found that interdisciplinarity is difficult to accomplish in a meaningful way given issues such as unfamiliarity and tensions between different disciplines.

Though the value of new forms of hackathon is represented by their increased use, commentators suggest that work is required in developing these approaches so that they are able to fulfil clearly defined aims, and harness interdisciplinary knowledge in meaningful ways (Taylor & Clarke 2018). Through our work we aim to contribute to this research, detailing some of the opportunities, challenges and implications of designing a non-conventional *ethical* hackathon.

1.2 Emergence of the Ethical Hackathon Approach

Over the last decade a programme of work - Responsible Innovation (RI) (Stilgoe et al. 2013) - has emerged as a concept and funding requirement to explore ways to integrate ethical and societal considerations into the processes and outcomes of research and innovation, to design technologies for greater societal benefit. A key

outcome of a specific investigation that sought to embed RI into ICT practices (Jirotka et al. 2017) was a twist on the idea of a hackathon. This approach includes an explicit RI element by requiring collaboration between researchers and practitioners from different disciplinary backgrounds and with potentially different concerns and values, to identify and address societal and ethical issues arising from ICT as a creative problem-solving activity. The name attributed to this approach was carefully crafted to incorporate its key commitments; however, a legal issue emerged as the name we selected had previously been claimed by a large international organisation. Thus we were required to adapt, and termed the approach `ethical hackathon'.

In an ethical hackathon, computer scientists and engineers collaborate with specialists from other fields to learn how to work together effectively to design more responsible technologies. Teams which include computer scientists, engineers, ethicists, social scientists and business students, complete a task that requires them to anticipate and reflect on the social and ethical issues that may emerge from an innovation, and also consider how to address these in their technical designs. Depending on the task, teams might produce a design document, mock up or a prototype. Entries are evaluated by judges in terms of how the ethical issues were identified and addressed, along with traditional hackathon parameters such as functionality, efficiency and safety. In this way the ethical hackathon enables RI issues to act as a resource for creativity in ICT innovation. This is important as an extended study conducted with key ICT researchers and practitioners in the UK found that many of the researchers were initially resistant to the idea of considering the potential challenging social and ethical consequences of their research (Eden et al. 2013). Towards the end of many interviews however, some acknowledged that they simply had never been asked to consider these kinds of questions. The ethical hackathon approach gives technical innovators the space for deliberation in collaboration with other stakeholders from different disciplines and institutions, to provide different perspectives on the intended and unintended consequences of an innovation and to suggest ways to address them. This approach has been tested and refined in various ways, some of which are described in this paper.

2. Methods: Ethical hackathon event

We ran an ethical hackathon event that challenged interdisciplinary student teams to identify and address ethical issues in the design of AI systems. This was part of the UnBias Project (https://unbias.wp.horizon.ac.uk) which explored the user experience of algorithm driven internet services and the processes of algorithm design. There was particular interest in circumstances in which algorithmic processes might produce bias or unfair outcomes. The mission of the project was to co-design alongside relevant stakeholders recommendations and materials for the design, regulation and education of algorithms to promote 'fairness' in their increasingly pervasive use in contemporary life. The event took place over a weekend in the summer of 2018, at a UK city. Participants were recruited via our project website and network, and relevant mailing lists. The context and purpose of the event was set out as follows:

Artificial Intelligence shapes digital services that have become central to our everyday lives. Online platforms leverage the power of AI to monetise our attention, with often unethical side-effects: our privacy is routinely breached, our perception of the world is seriously distorted, and we are left with unhealthy addictions to our screens and devices. The deep asymmetry of power between users and service providers, the opacity and unaccountability of the algorithms driving these services, and their exploitation by trolls, bullies and propagandists are serious threats to our well-being in the digital era.

This hackathon invites participants to build tools to empower users in their online lives. The tools might address a relevant problem in this space, including (but not limited to) filter bubbles and fake news, biased and unaccountable algorithms, or the profit-driven metrics that guide these AI-powered services.

The task was formulated to be open so that participants could pursue their own areas of interest. Approximately 50 people applied to take part in the event and 30 attended - a fairly typical drop-off rate for a free event. With the exception of one professional programmer, the participants were all undergraduate or postgraduate students; their topics of study were: Computer Science (7); Artificial Intelligence (5); Innovation, Technology and Law (3); Digital Society (2); Data Science (1); Film Directing (1); Economics and Media Sciences (1); Accounting and Finance (1); Cognitive Science (1); and others (8). We were therefore able to achieve a wide interdisciplinary mix.

The event was organised and facilitated by members of the project team and an events management company. The schedule was balanced between structured sessions and time for the teams to work independently. The start of the first day focused on participant networking and team building. Three educational talks were also given to help raise participants' awareness of ethical issues in AI. These covered ethics in the digital age, issues around the purposefully addictive design of user interfaces and the General Data Protection Regulation (GDPR). Once the teams were formed they discussed ideas and agreed on a tool to develop. For guidance they were shown the judging criteria below. These criteria were designed to encourage attention to both ethical and technical issues.

- Relevance: does the work address an important / relevant ethical problem?
- Grounding: Is the proposed solution solidly grounded in ethical principles?
- Business case / sustainability: Would this system be economically viable?
- Design and user experience: Is the user interface well designed?
- Technical excellence: How close is this to being implemented?

Given the openness of the task, teams were not given specific data sets to work with, in contrast to other hackathon events. Teams were encouraged to identify open and ethical data sources and given assistance with this where necessary. The teams had regular feedback sessions with event mentors and could also request help at any time. Nine mentors attended over the course of the two days. They consisted of three academics, four professional programmers and two data scientists with mixed expertise in AI, software development, computer ethics and RI. The mentors tried to ensure that teams stayed on task and addressed all the judging criteria. The teams spent between five to eight hours working independently on their designs on day one and up to four hours on day two. At the end of day two each team gave a presentation of their work to the other teams and a judging panel. The judges were selected to represent different sectors: academia, industry and art/design. Prizes were awarded to the teams judged as best meeting the assessment criteria. The winning projects were:

- 1st prize: a tool to enable large companies to analyse their past hiring practices and identify possible discrimination against employees with protected characteristics etc.
- 2nd prize: a Facebook plug-in to reduce the addictive aspect of its user interface.
- 3rd prize: a tool to check whether websites are GDPR compliant.

In order to assess the ethical hackathon and its outcomes, participants were asked to fill out questionnaires. At the start of the event they were asked about their previous experiences in dealing with ethical issues arising from technology and their expectations about the event. At the end of the event they were asked to rate their experiences and reflect on what they had learnt. We carried out brief (10-15 minute) interviews with event participants and collected our own fieldwork observations. Our project team had a clear set of objectives for the event; namely to accomplish:

- The integration of ethics into an extension of the conventional hackathon model.
- The treatment of ethics as equal to technical issues in the task set for participants.
- Interdisciplinary teamwork and collaboration in order to facilitate shared insights and peer learning.

Analysis of our data enabled us to identify both opportunities and challenges associated with meeting these aims.

3. Findings

Overall we found the ethical hackathon approach to be successful. The event was well attended with, as described above, a wide range of disciplines represented. Thirteen of the participants completed a questionnaire at the start of the event, of whom six had never attended a hackathon. Only six of the respondents said they had taken ethics modules as part of their studies. We therefore felt satisfied that our event had successfully targeted a mixed audience. Eight participants completed a questionnaire at the end of the weekend. Of these, seven rated their enjoyment of the event, on a scale of 1 (worst) to 10 (best), as a 7 or above. All eight rated their level of interest as a 7 or above and their scores for relevance to their personal studies/work ranged from 5 to 10. Analysis of our qualitative data from free text questionnaire responses, interview responses and fieldwork observations enabled a more detailed examination of how well the event

ran and the outcomes it produced. We identified three particular issues that represent challenges to running an interdisciplinary ethical hackathon. These relate to:

- The framing of ethics
- Engendering interdisciplinary teamwork and meaningful collaboration
- Varying familiarity with the hackathon model

These challenges are discussed below through the use of exemplary quotes from participants and our observational reflections. We demonstrate that this novel hackathon model can offer a valuable approach towards meaningful interdisciplinary collaboration through an RI lens; however, this is not a simple endeavour and requires extremely careful planning to anchor it to its key goals.

3.1 Framing Ethics

Ethical considerations were forefronted in our recruitment material, the educational talks and the challenge set for the task. We observed numerous instances of attendees engaging with ethical aspects of their design - interacting with one another and mentors to explore their perspectives. These occurred over the entire event but particularly when teams were deciding upon a tool to develop. The teams discussed the ethics of AI in a diverse range of domain areas. Our analysis indicates that the experience enabled participants to deepen their awareness and understanding of ethical issues in useful ways. Both our questionnaires asked participants to rate their level of confidence (from 1 to 10) to embed ethical issues in their own work. Of the five participants who completed the opening day and end of event questionnaires, four gave a confidence level in the second questionnaire that was higher than the first and their scores in the second questionnaire ranged from 7 to 10. The following exemplary quote (Example 1, below) from an interview respondent characterises how participants from different backgrounds dealt with ethics in design as a procedural matter during the event:

Example 1: [law student, interview]
People with more [different] backgrounds, they don't focus on ethical aspects in advance, whereas lawyers, because it's how law is created, basically it starts from ethics and then it's the law. So, I don't know if I learnt anything about ethics, but I would say I learnt about how everybody thinks about ethics or how they approach such topics.

However, despite this articulation of engagement with ethics, we largely observed that teams did not appear to forefront ethics in relation to the design of their tools in the integral manner we desired - through an RI lens. Once decisions had been made about tool development, ethics was often treated as of less importance than technical considerations. As indicated in Example 2 (below), participants tended to focus on producing a tangible outcome to the challenge. Even when prompted to do so by mentors, the teams often only appeared to pay brief attention to embedding ethical considerations into the processes of their tool development. Instead, primacy was mostly given to technical work, as posited by the respondents in Examples 2 and 3 (below):

Example 2: [social science student, interview]
In my group, they ... made a bad Twitter account to calculate activity of the comments. At first, I have some concerns about whether we have the right to [use] the Twitter data because there is terms and conditions for Twitter that we actually cannot access to the private - personal posts. I have that concern in mind, but I didn't bring it up, because I think if that is the case then we cannot do that project. So, it's quite embarrassing.

Example 3: [social science student, interview]
I think they are more focused on the programming because [one team member] said, I just love programming. I don't want to think about other things.

Example 4 (below), is a statement from a technical participant who seems to reinforce this issue by conflating ethics with `business planning'. This demonstrates what appears to be a lack of engagement with the importance of ethics in its own right and as of equal importance to the technical aspects of the task:

Example 4: [professional programmer, questionnaire part 2]
[on rating interest in the event at 7/10] *Very interested in the task, less interested in business planning*

Though there were positive aspects to considering ethics, on the whole we found that as the teams proceeded with their work a hierarchy emerged in which ethics was treated as lesser component of design than technical issues - a finding we will revisit in the discussion section of the paper.

3.2 Interdisciplinary Teamwork and Meaningful Collaboration

Bringing together participants from disciplines who may not be accustomed to working together, was essential to the ethical hackathon. We viewed interdisciplinarity as fundamental to responding to the complex issues related to the use and development of AI and we also received a great deal of positive feedback about it from participants –-- see Example 5 (below). The structure of the event also aided interdisciplinarity: its relatively short time scale prevented groups from attempting to fully utilise AI, which requires a great deal of technical expertise, when responding to the challenge. To some extent this helped prevent a separation of technical participants from those with non-technical expertise, also providing a means for teams to work on the ethics of AI without undertaking the complex work of *doing* AI.

Example 5: [social science background, interview]
I think the most important thing is to cooperate. The cooperation between humanist and the computer scientist. So, I think this environment is really good to bring these two together and - because now there are so many issues concerned with the technology and I think sociologists have their role, to improve all this process.

The participant quoted in Example 6 (below) also comments positively on this interdisciplinarity. However, he also touches upon some of the more problematic interactions we observed between technical and non-technical participants.

Example 6: [computer science background, interview]
We didn't get to do a lot of technical stuff, but as I said earlier to you, it's actually a nice change… I think if it was just informatics people, we would be certainly further on in the actual development coding-wise, but we would have completely neglected the business side of things and what problem are we actually trying to solve, because that's not something we usually really think about.

On the whole there seemed to be divisions that emerged between those with technical expertise and those from other disciplines. The participant in Example 6 characterises this as a falling behind in the development of the code given the presence of non-technical participants. In Examples 7 and 8 (below) non-technical participants characterise their own contributions as negative - describing their roles as concerning `background issues' or themselves as the `weak link'.

Example 7: [law student, interview]
I feel like my help so far has been more like with background issues more than leading. So, for example when we were discussing which topic we can do, I told them about this issue that ethnic names are chosen less in a job than white names... So, I give them those ideas and then we can start talking about it, but I'm not sure how much legal help it has been.

Example 8: [law student, interview]
I got the feeling that I was the weak link in the group, like `how can I help you?' I was giving ideas. I was like: `can we build that? Is that possible?'

These comments correspond with our own observations. Technical participants typically led the teams and steered the tool development. Divisions between technical and non-technical participants emerged; subgroups often formed in which technical participants undertook the bulk of development work including coding etc. whilst non-technical participants were relegated to less useful tasks such as creating logos. There were at times long periods with no direct interaction between sub-groups and little emphasis placed by teams on what their different disciplines were bringing to the task. Indeed, challenges in communication appeared to be common, as illustrated in Example 9 (below):

Example 9: [social science background, interview]

I find it really difficult to communicate with - those three guys in my group are classmates... So, they are talking about all those technical things and I don't really understand sometimes. But only when it is related to the issues themselves, I can have some comments or have some feelings, so I find it quite difficult.

As the participant suggests, non-technical participants may often have felt excluded from technical discussions as they lacked relevant expertise and vocabulary.

Though there were some positive responses to the interdisciplinary nature of the event, technical knowledge was often treated as superior to that arising from other disciplines. It seemed to be rare that detailed explanations of disciplinary areas such as law or the social sciences were solicited from non-technical team members by their technical counterparts. This seemingly limited mutual knowledge exchange and learning, demonstrates that solutions may need to be found to foster genuine collaboration between technical and non-technical disciplines.

3.3 Varying Familiarity with Hackathons

As noted above, the interdisciplinary nature of the event meant participants had varying degrees of familiarity with the notion of a hackathon. Some (technical) participants reported taking part in up to seven hackathon events previously whereas for others (in particular non-technical participants) this was their first such event. This created different kinds of challenges. Whilst first-time hackathon participants expressed excitement at the opportunity to take part in the event, they also reported, as the law student in Example 10 (below) suggests, not being clear about how the event would unfold:

Example 10: [law student, interview]
When we made the groups and then we were taken to the room, what now? Are they going to give us instructions? ... So I think I was waiting for more instructions. I don't know. I was a little bit confused of how it works, but it was good.

By contrast, those experienced with hackathons - e.g., Example 11 (below) - often assumed that our event would take a more conventional trajectory and did not anticipate the differences in format and task that occurred:

Example 11: [computer science background, interview]
So, it's not like your traditional hackathon, where you have a full demo at the end necessarily. It might be good to make that a little bit more clear at the start...We were very worried that we wouldn't have something in the end to represent. Then we were looking at the judging sheet and it was like, okay, it's not just about that.

Differing levels of familiarity with hackathons may have created a misalignment of expectations and contributed to the hierarchy observed above where experienced technical participants were deferred to as experts and guides for the others. In turn this also appeared to contribute to the precedence given to technical concerns over ethical ones. Varying familiarity with hackathons therefore presents a challenge to running events that successfully promote interdisciplinarity in the ethical design of AI systems.

4. Discussion

The ethical hackathon is designed to stimulate participants to anticipate ethical and societal consequences of innovation might be from the outset of the innovation cycle- to deliberate on the issues amongst themselves as representatives of different stakeholders and to attempt to mitigate issues that may arise. Our approach emerged from the field of Responsible Innovation, and extends the traditional hackathon model to give equal importance to ethical considerations alongside technical development. We conceptualise the ethical hackathon as a space for the harnessing of interdisciplinary knowledge to address ethical issues. We have analysed data collected at a recent ethical hackathon event to evaluate the opportunities and challenges of this approach.

Given the complexities of bringing together different disciplines, we regard our event as a success. It was well attended, with enthusiastic participants from a range of disciplinary backgrounds. We received positive feedback and observed many instances of interdisciplinary collaboration and peer learning. However, we also

identified that there may be challenges in hosting a truly interdisciplinary event where i) ethics is fully integrated and treated equally to technical issues and where ii) there is meaningful collaboration between team-members from different disciplines. In particular, it may be that design of our event unintentionally perpetuated existing disciplinary barriers between those with and without technical expertise, and in turn appeared to present a boundary to the appropriate integration of ethical considerations. Though ethical considerations were forefronted from the recruitment phase of the ethical hackathon, we observed that our participants tended to treat ethics as a secondary consideration to technology development.

The varying manner in which participants from different disciplines responded to the challenge indicates potential issues with the framing of the event. Various issues may have contributed to misaligned expectations. There was no explicit demonstration of how ethics should be applied to the design process through the lens of RI. Instead, participants were informed about the centrality of ethical issues to AI and the design challenge, then decided themselves how to incorporate ethical considerations in the development of their tools. Although mentors often advised teams to have a stronger integration of ethics in their work, this was at times not fully taken on board. Instead, the ethical concerns largely seemed to be side-lined in relation to the technical development of the tool. Additionally, the use of the term `ethical hackathon' - as noted earlier in the paper a required change from the name originally planned for this kind of event - might have added to this problem. The term ethical hackathon not only gives the misleading impression that ethics can be hacked, but it also appears to elevate the role of those who have familiarity with hackathons and who also have technical expertise. In our event, those less familiar or without experience of hackathons looked to their more experienced counterparts as `expert', thus the disciplinary hierarchy was reinforced. As we saw in the findings, this appeared to leave many without such experience to feel inferior or excluded. Finally, the task itself may also have played a part in reinforcing disciplinary boundaries. Though the description of the task was left open, solutions which relied on technical knowledge still dominated - creating divisions between what different team-members were able to contribute.

Despite the success of our event, it appears to be the case is that a misalignment of expectations may have limited the emergence of a *truly* interdisciplinary ethical hackathon. From these lessons learned, we now suggest implications for design of these events in the future:

- Renaming the `ethical hackathon': we suggest that the name `RI-athon' may be more appropriate. This forefronts the importance of responsible innovation - namely identifying and addressing ethical concerns of an innovation - and removes the problematic dominant technical connotations associated with the term `hackathon'.
- Framing ethics: activities that, through the lens of RI, enable participants to engage with and experience *how* ethical considerations can be applied to their designs are essential. We suggest that relevant presentations and also interactive tasks such as short case study-based exercises-are important to undertake before presenting the main challenge.
- Setting an inclusive challenge: it is important that the design challenge is inclusive and accessible to all team members, and so does not require an intricate technical outcome. Technical output could be based on 'lightweight' outcomes such as conceptual prototypes and mock ups etc.; this would avoid long periods of separation between non-technical and technical areas and allow for more consistently interdisciplinary group input.

5. Future Work and Conclusions

We will apply the lessons learnt through the conduct of more events - perhaps more appropriately called `RI-athons'. We will also carry out deeper evaluations of how stakeholders interact with the multidisciplinary nature of the event, and how far the goals of an RI-athon are reached. Importantly we recognise that an iterative approach, where each event will develop and build on the last will be required given the complexities that are involved. Though this is not a simple endeavour, the importance of an interdisciplinary approach to address concerns over innovative technologies such as AI makes it a necessary one.

6. Acknowledgements

We would like to thank all the participants who gave their time to take part in the ethical hackathon and associated data collection activities. This research was conducted as part of the project ‘UnBias: Emancipating

Users against Algorithmic Biases for a Trusted Digital Economy'. This project was funded by the UK's Engineering and Physical Sciences Research Council (EPSRC). Project reference EP/N02785X/1.

7. References

ACM (2018), 'Deconstructing the ACM Code of Ethics and Professional Conduct', [online], *http://doi.acm.org/10.1145/306286.306291*

AI Matters (2017), 'New Conference: AAAI/ACM Conference on AI, Ethics, and Society', [online], *https://sigai.acm.org/aimatters/blog/2017/09/20/new- conference-aaaiacm-conference-on-ai-ethics-and-society/*

Albrecht, S. V., Beck, J. C., Buckeridge, D. L., Botea, A., Caragea, C., Chi, C.-h., Damoulas, T., Dilkina, B., Eaton, E., Fazli, P., Ganzfried, S., Giles, C. L., Guillet, S., Holte, R., Hutter, F., Koch, T., Leonetti, M., Lindauer, M., Machado, M. C., Malitsky, Y., Marcus, G., Meijer, S., Rossi, F., Shaban-Nejad, A., Thiebaux, S., Veloso, M., Walsh, T., Wang, C., Zhang, J. & Zheng, Y. (2015), 'Reports on the 2015 AAAI Workshop Program', *AI Magazine* 36(2), 90, [online], *https://aaai.org/ojs/index.php/aimagazine/article/view/2590*

Birbeck, N., Lawson, S., Morrissey, K., Rapley, T. & Olivier, P. (2017), Self Harmony: rethinking hackathons to design and critique digital technologies for those affected by self-harm, *in* 'Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems', ACM, pp. 146– 157.

Boyles, J. L. (2017), 'Laboratories for news? Experimenting with journalism hackathons', *Journalism* p. 1464884917737213.

Briscoe, G. (2014), 'Digital innovation: The hackathon phenomenon', *Working Papers of The Sustainable Society Network* .

Cutler, A., Pribic´, M. & Humphrey, L. (2018), Everyday Ethics for Artificial Intelligence, Technical report, IBM.

Eden, G., Jirotka, M. & Stahl, B. (2013), Responsible re- search and innovation: Critical reflection into the potential social consequences of ICT, *in* 'Research Challenges in Information Science (RCIS), 2013 IEEE Seventh Inter- national Conference on', IEEE, pp. 1–12.

Edmondson, A. C. & Harvey, J. F. (2018), 'Cross-boundary teaming for innovation: Integrating research on teams and knowledge in organizations', *Human Resource Management Review* 28(4), 347–360.

Floridi, L. (2010), *The Cambridge Handbook of Information and Computer Ethics*, Cambridge University Press.

IEEE (2018), 'Ethics and Member Conduct', [online], *https://www.ieee.org/about/ethics/index.html*

Jirotka, M., Grimpe, B., Stahl, B., Eden, G. & Hartswood, M. (2017), 'Responsible research and innovation in the digital age', *Communications of the ACM* 60(5), 62–68.

Johnson, D. G. & Miller, K. W. (2009), *Computers Ethics*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Porter, E., Bopp, C., Gerber, E. & Voida, A. (2017), Reappropriating hackathons: the production work of the CHI4Good day of service, *in* 'Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems', ACM, pp. 810–814.

Society, B. C. (2015), 'Code of Conduct for BCS Members', [online], *https://www.bcs.org/upload/pdf/conduct.pdf*

Stilgoe, J., Owen, R. & Macnaghten, P. (2013), 'Developing a framework for responsible innovation', *Research Policy* 42(9), 1568–1580.

Taylor, N. & Clarke, L. (2018), Everybody's Hacking: Participation and the Mainstreaming of Hackathons, *in* 'Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems', ACM, p. 172.

Taylor, N., Clarke, L., Skelly, M. & Nevay, S. (2018), Strategies for Engaging Communities in Creating Physical Civic Technologies, *in* 'Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems', ACM, p. 507.